III. Linkage analysis

- Scope and concepts
- Parametric and nonparametric methods
- Issues: reduced penetrance, phenocopy, heterogeneity
- Practice: LINKAGE, GENEHUNTER, Merlin, SOLAR, SIMULATE and SLINK

Scope

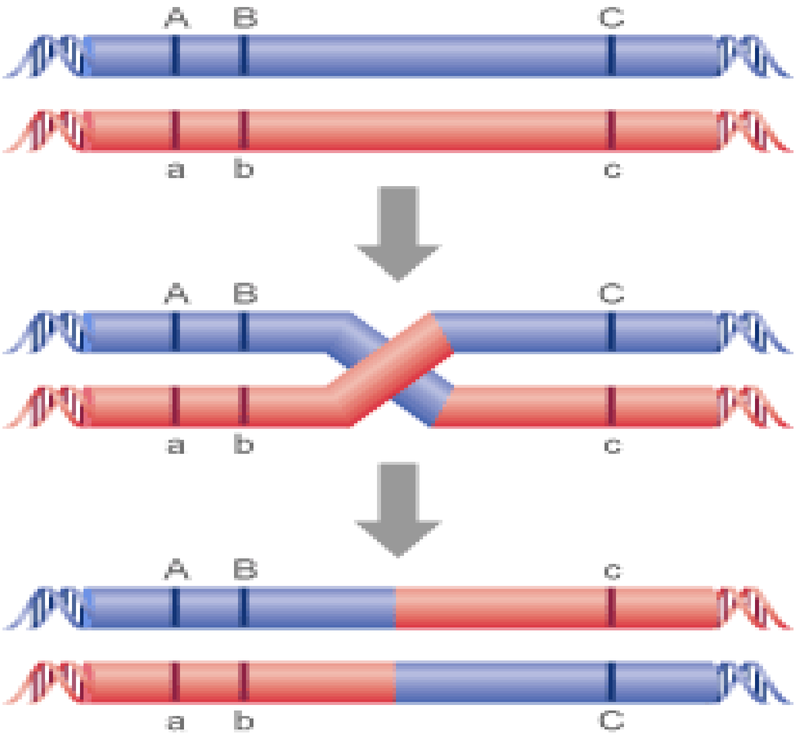
- To study frequency of cosegregation of trait and alleles at putative marker loci in families.
- It is desirable since for disease with low penetrance, it is more likely familial aggregation would be more genetic than environmental; data are likely to be more accessible and thus have played a key role in positional cloning.
- Parametric (model-based) method refers to the situation when the mode of inheritance (allele frequency, penetrance) is fully specified, otherwise it is nonparametric (model-free).
- The statistical model is encapsulated in segregation analysis but with marker information, map function is also relevant. Moreover, methods for detecting marker-disease association through linkage disequilibrium are special cases of model-based linkage analysis.

Reminders

- **Genetic linkage** refers to the phenomenon that loci on the same chromosome tend to segregation together during meiosis.
- Because of **crossovers**, alleles on the same chromosome can be separated and go to different daughter cells with the probability being greater if the loci involved are far apart.
- The frequency that a crossover occurs is call **recombination rate** (θ). θ=0.5 indicate independence, i.e., an equal chance of linkage vs no linkage. A centimorgan (cM) indicates that the frequency of recombination is 1%. A **genetic map** catalogues the recombination rates between a set of genetic loci.

A schematic sketch of recombination

-



Crossover process

- The simplest case assumes no interference so crossovers occur as a Poisson process; in which case map function M(x) is obtained from the probability that a Poisson r.v. with mean x is odd,

$$\begin{aligned}\textstyle\sum_{k=odd}\exp(-x)x^k/k! &= \exp(-x)/2\textstyle\sum_k\left(x^k/k! - (-x)^k/k!\right)\\ &= (1-\exp(-2x))/2\end{aligned}$$

- Alternatively, $M(x+\delta x) = M(x)(1-\delta x) + (1-M(x))\delta x$

$$dM(x)/dx = (1-2M(x))$$

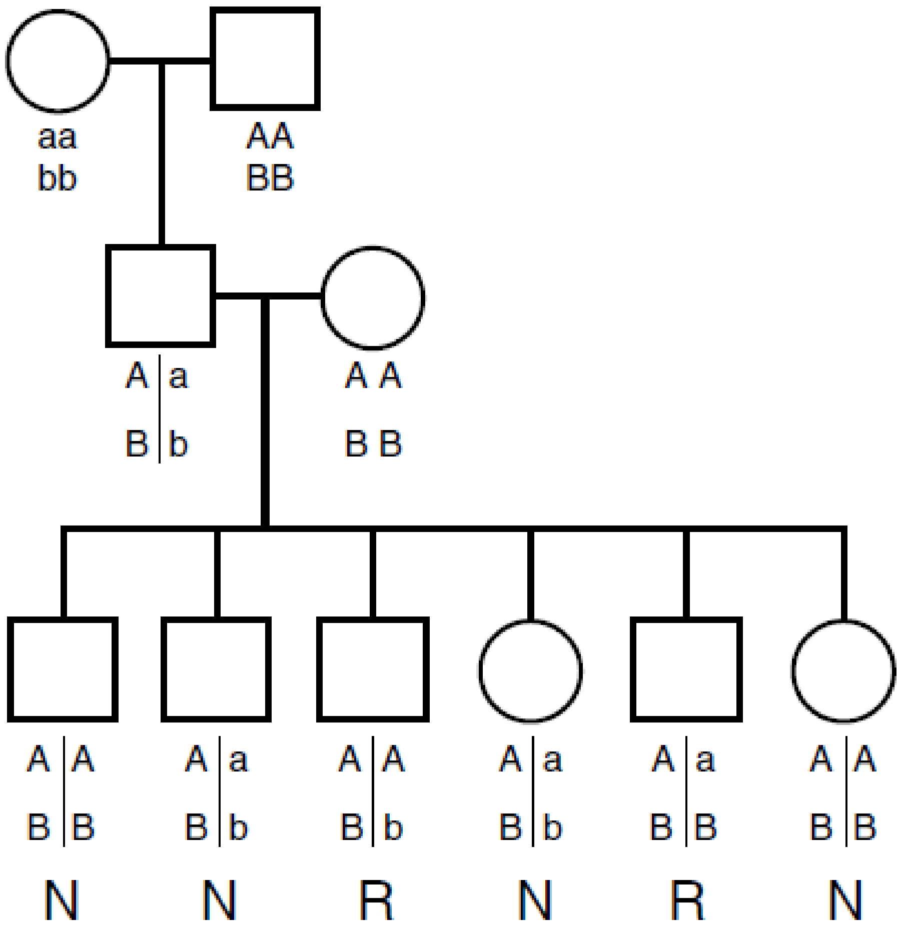
$$\ln(1-2M(x)) = -2x$$

Lod score method

- When recombinants (R) and non-recombinants (NR) can be observed from families, the likelihood is a binomial function the recombination, $L(\theta)=(1-\theta)^{N-R}\ \theta^{R}$.
- The lod-score is defined as $Z(\theta)=\log_{10}(L(\theta)/L(0.5))$, therefore a lod score of 3 is commonly used to indicate linkage, and -2 to indicate no linkage. In the former case, the probability of linkage vs no linkage is 1000:1.
- Note the close relationship with log-likelihood ratio test (natural log of the probability ratio, or **location score**). More specifically, $4.6\text{LOD}\sim\chi_1^2$, a chi-squared distribution with one degree of freedom.
- The test is one-tailed and pointwise significance determined by

$$0.5\ \ (\chi^2\ >\ 2\ln 10\ \ \text{lod})$$

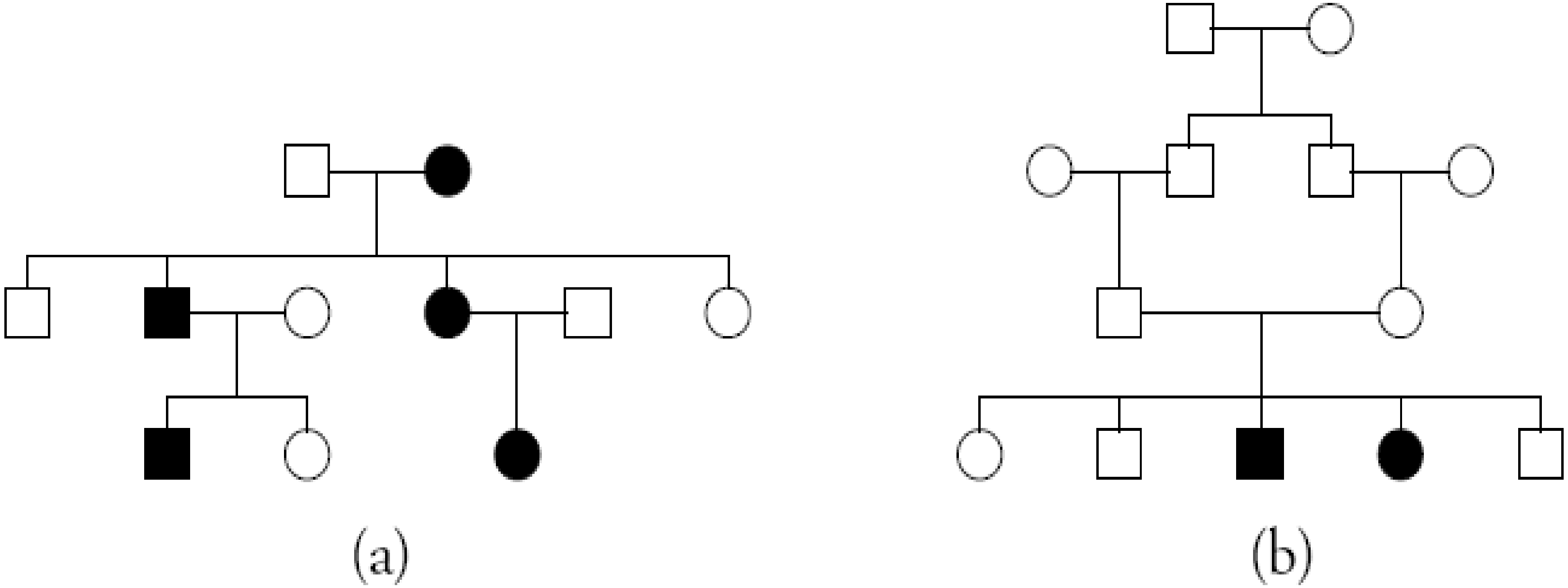
- since $2\ln10\ Z(\theta)$ is a 50:50 mixture of a point mass at 0 and χ_1^2.



$$Z(\theta) = \log_{10}\frac{\theta^2(1-\theta)^4}{(0.5)^2(0.5)^4} = \log_{10} 2^6\theta^2(1-\theta)^4$$

A phase-known pedigree with two codominant loci with alleles A/a and B/b. Nonrecombinant and recombinant meioses are marked with N and R, respectively. The MLE of recombination rate (θ) is 2/6 and lod score 0.1475. For phase-unknown pedigree(s), one need to consider all possible phases and parameter estimation is customarily done through maximum likelihood

Disease models



- Two pedigrees with typical (a) autosomal dominant inheritance and (b) autosomal recessive inheritance. Affected individuals have filled symbols

Lander-Green algorithm

$$\begin{aligned} p(a,b) &= \sum_{\alpha,\beta} p(a,b|\alpha,\beta)p(\alpha,\beta) \\ &= \sum_{\alpha,\beta} p(\alpha,\beta)p(a|\alpha)p(b|\beta) \\ &= \sum_{\alpha} p(\alpha)p(a|\alpha)\sum_{\beta} p(\beta|\alpha)p(b|\beta) \\ &= 2^{-N}P_aT_1P_b \end{aligned}$$

- For loci A and B, the probability of observing phenotype data a and b can be expressed as follows (Idury & Elston, 1997), with $T_1=2^N x 2^N$ transmission matrix with elements $t_{\alpha,\beta}=\theta^p(1-\theta)^{N-p}$, where θ is the recombination rate, p determined by α,β, the inheritance vectors at A and B.

Complex disease

- For complex traits such as schizophrenia, a small amount of phenocopy, reduced penetrances in a variety of liability classes are customarily defined according to age groups.
- It is usually to conduct simulation study to examine power of given family structures and disease models, via computer programs such as SLINK. The behavior of the linkage statistics under the null hypothesis can also be assessed.
- The typical simulations also allow for linkage heterogeneity such that linkage occurs in some but not all families.
- However, as will be seen it is also preferable to avoid specification of such models through nonparametric linkage analysis.

Genetic heterogeneity

- Assuming that while difference pedigrees may have different genetic forms of the disease, within a pedigree only a single genetic form is present, it leads to an admixture model in which the probability of the trait being linked and unlinked in a given pedigree is α and 1-α, respectively.
- Therefore two different lod scores are calculated: the standard lod score which assumes genetic homogeneity, and a lod score which allows for maximization of the likelihood function over both the recombination rate and linked fraction α.

Generalized linkage statistics

- A set of model-free linkage statistics can be defined by fixing the prevalence of disease in a population, dominant and recessive models with reduced penetrance
- MALOD – the maximum admixture lod score (over transmission models and α)
- MFLOD – the lod score obtained from the difference between likelihood maximizing over penetrance and α, and likelihood maximizing over penetrance but setting α=0.
- They appear to be similar for small families, suggesting that uncertain mode of inheritance is not a serious issue for linkage analysis of genes with minor effects. However, it has been pointed that nonparametric linkage (NPL) statistics could be conservative between markers.

- (Curtis and Sham AJHG, 1995; Sham et al. AJHG, 2000)

Power of linkage study

- We start from the simple case of direct counting is possible.
- For a true recombination fraction, θ, significance level α, and power 1-β, the required sample size is approximately,

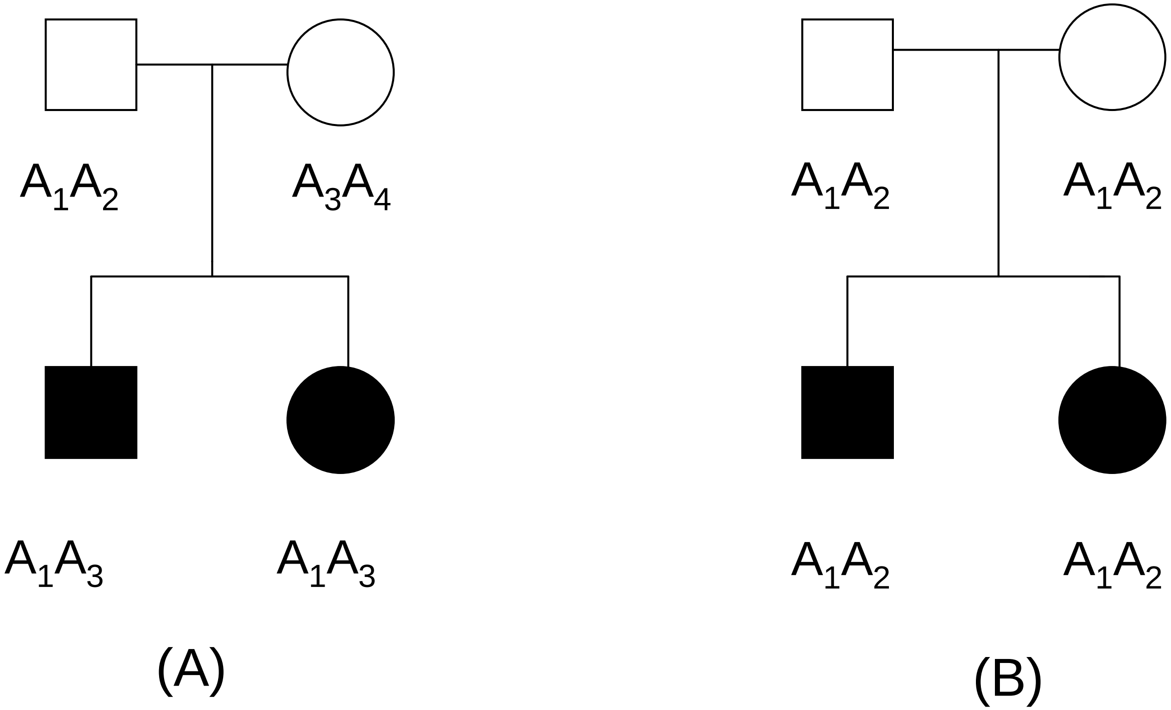
$$N = \left[(z_{1-\alpha}/2 + z_{1-\beta}\sqrt{\theta(1-\theta))}/(\theta - 1/2)\right]^2$$

- For z_α=3, 1-β=0.8, θ=0.05, we have N=20, the total number of recombinants and nonrecombinants.
- A good picture of the properties of the disease include its phenotype, mode of inheritance, population frequency. One offspring is added to each sibship for phase-unknown, and adjust for incomplete marker informativeness by inverse heterozygosity, for incomplete penetrance by inverse expected lod score, for heterogeneity by relative efficiency, for misclassification, etc.

Computer simulation in linkage analysis

- The major uses of computer simulation in linkage analysis are to estimate power, obtain empirical p-value, and explore newly-proposed methods. The idea is to generate genotype or phenotype data using random numbers, conditional on given pedigree structures, under a specified genetic model and recombination fraction. Simulation involves generating multiple sets of such data and analyzing them using a linkage program. The following summary statistics are often of interest:
  - P-value: the probability that the lod score will exceed the critical value (usually set at 3) under no linkage.
  - Power: the probability that the lod score will exceed the critical value (usually set at 3) under linkage.
  - ELOD: the expected value of the lod score at the true recombination fraction

Nonparametric linkage



- (A) Identity-by-descent (IBD) if sharing the same allele of a ancestor
- (B) Identity-by-state (IBS) if sharing the same allele, regardless the ancestral origin

Affected sib-pair (ASP) method

$$P(IBD = 0) = \frac{1}{4} - \frac{(\Psi - 0.5)V_A + (2\Psi - \Psi^2 - 0.75)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

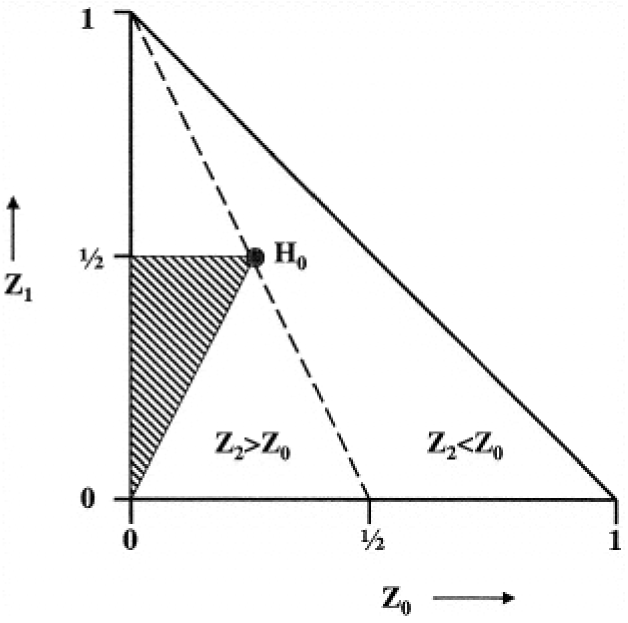
$$P(IBD = 1) = \frac{1}{2} - \frac{2(\Psi^2 - \Psi + 0.25)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

$$P(IBD = 2) = \frac{1}{4} + \frac{(\Psi - 0.5)V_A + (\Psi^2 - 0.25)V_D}{4(K^2 + 0.5V_A + 0.25V_D)}$$

- Therefore under the null hypothesis of no linkage, pairs of sibs are expected to share 0, 1, 2 with probabilities ¼, ½, ¼, respectively.
- When there is linkage between the putative disease locus and marker, one may see allele-sharing between affected sib pairs with higher probabilities. The contrast between the observed and expected frequencies can be used as evidence of linkage.

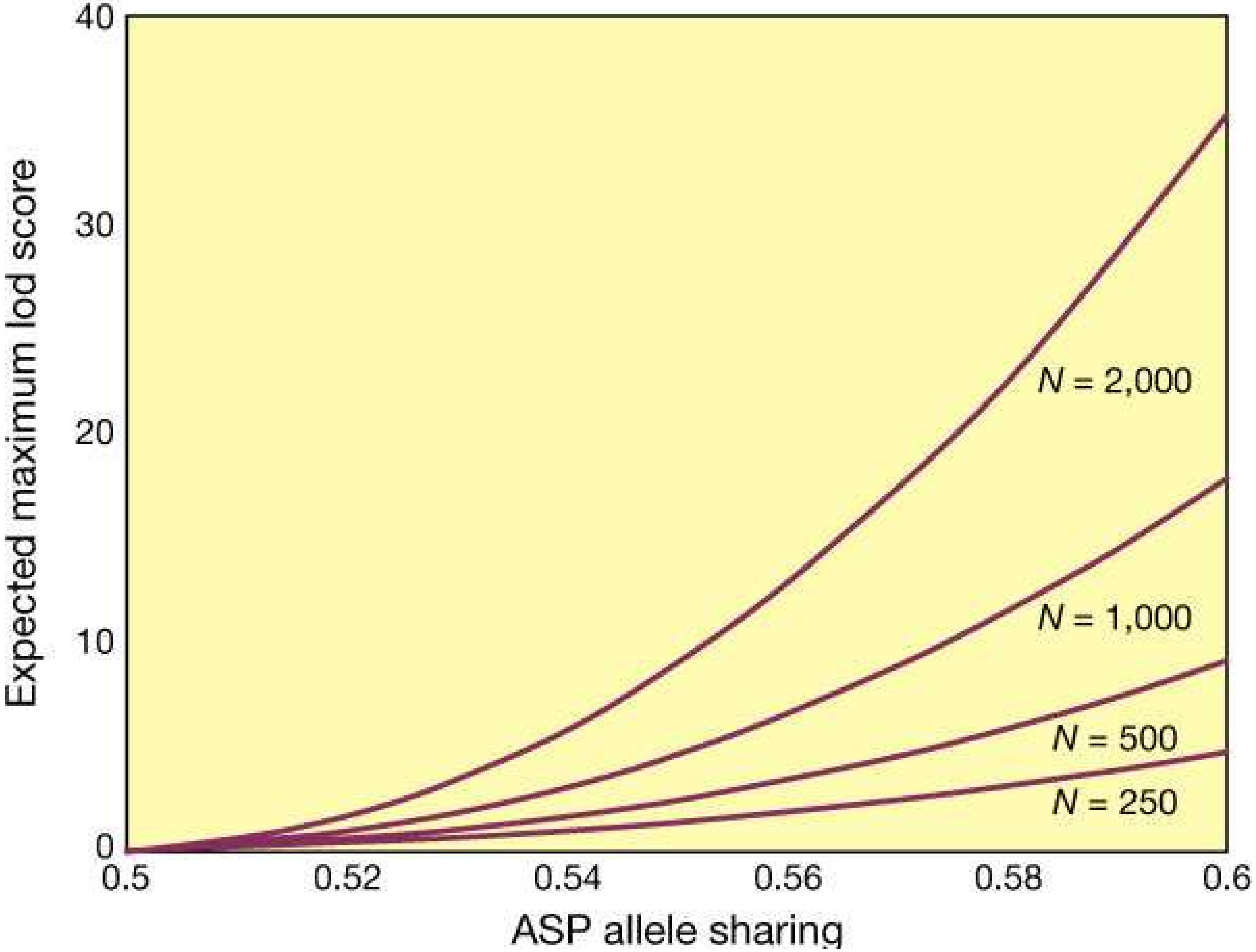
Tests using ASPs

- Simple χ^2 test $S_1 = \sum_i (o_i - e_i)^2 \Big/ e_i$
- Means test $S_2 = \left(\frac{1}{2} n_1 + n_2\right) \Big/ \sqrt{\frac{n}{8}} \sim N(0,1)$
- MLS $L = \sum_i z_i P(x \mid z_i)$

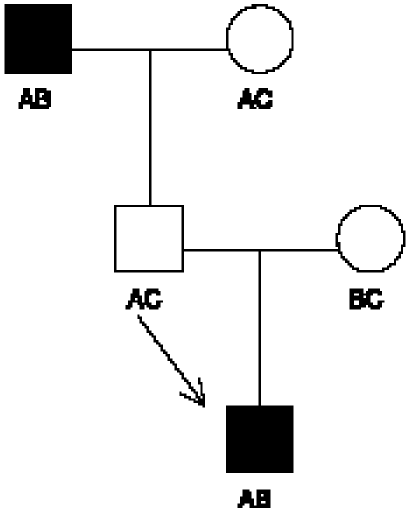


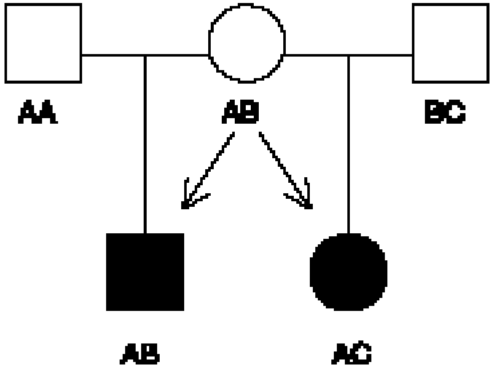
- Sham & Zhao (1998) Sham & Zhao (1998) in Bishop MJ (Ed) Guide To Human Genome Computing, 2nd Edition
- Elston (2001) AJHG 69: 1149-50

ASPs required to detect linkage as a function of allele sharing (Risch (2000) Nature)



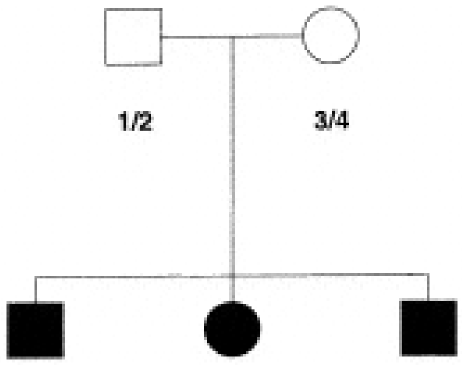
Affected relatives method





- The idea of affected sib pairs can be extended to affected relatives.
- One can base on IBS, especially when highly polymorphic microsatellite markers are involved, e.g., affected pedigree member (APM) method by Weeks and Lange.
- It is more preferable to use IBD than IBS. Method to score allele-sharing between pairs of relatives have been developed by Whittemore and Halpern and implemented in GENEHUNTER as nonparametric linkage statistics (NPLpair, NPLall).

Three Affected Siblings (Nyholt (2000) AJHG)



Possible Genotype Configurations (n=20):

1/3	1/3	1/3
1/3	1/3	1/4
1/3	1/3	2/3
1/3	1/3	2/4
1/3	1/4	1/4
1/3	1/4	2/3
1/3	1/4	2/4
1/3	2/3	2/3
1/3	2/3	2/4
1/3	2/4	2/4
1/4	1/4	1/4
1/4	1/4	2/3
1/4	1/4	2/4
1/4	2/3	2/3
1/4	2/3	2/4
1/4	2/4	2/4
2/3	2/3	2/3
2/3	2/3	2/4
2/3	2/4	2/4
2/4	2/4	2/4

Inheritance Vectors (n=64):

1: 000000
2: 000001
010000
000100
3: 000010
100000
001000
4: 000011
001100
110000
5: 000101
010100
010001
6: 000110
011000
010010
001001
100100
100001
7: 000111
011100
010011
001101
110100
110001
8: 001010
101000
100010
9: 001011
101100
100011
001110
111000
110010
10: 001111
111100
110011
11: 010101
12: 010110
011001
100101
13: 010111
011101
110101
14: 011010
101001
100110
15: 011011
101101
110110
011110
100111
111001
16: 011111
111101
110111
17: 101010
18: 101011
101110
111010
19: 101111
111110
111011
20: 111111

Equivalent Inheritance Vectors:

1,11,17,20
2,3,5,8,13,16,18,19
4,10,12,14
6,7,9,15

Pair IBD Sharing:

IBD=2,2,2	4/64
IBD=2,1,1	24/64
IBD=2,0,0	12/64
IBD=1,1,0	24/64

NPL statistics

$$S_{pair} = 2/[n_A(n_A-1)] \sum_{1\leq k\leq l\leq n_A} \left[\frac{1}{4}\sum_{a=1}^{2}\sum_{b=1}^{2}\delta(s_{ka}, s_{lb})\right]$$

$$S_{all} = 2^{-n_A}\sum_{h}\left[\Pi_{i=1}^{2f} b_i(h)!\right]$$

$$Z = \sum_{i=1}^{m} Z_i/\sqrt{m},\ Z_i = (S_i-\mu_i)/\sigma_i$$

- where m=total # of pedigrees, n_A=# of affected relatives in a pedigree, 2f=total # of founder alleles, h=alleles from taking one allele from each affected individuals, b(h)=total # of appearances of a founder allele in the collect h, δ(.,.)=Kronecker delta function, s=inheritance vector.
- NPL_{pair} and NPL_{all} and are based on normal approximations of score functions S_{pair} and S_{all} and # of pairs of alleles from distinct pedigree members are IBD, and average # permutations that preserve a collection h, respectively.

IBS in affected relatives methods

- It can be sensitive to misspecification of marker allele frequencies.
- The null distribution of the test statistic may be skewed, leading to a potential anti-conservative test.
- It ignores IBD information available and less powerful than IBD-based method.

Parametric Linkage Analysis Using Extended Pedigrees (Gusella et al. (1983). Nature)

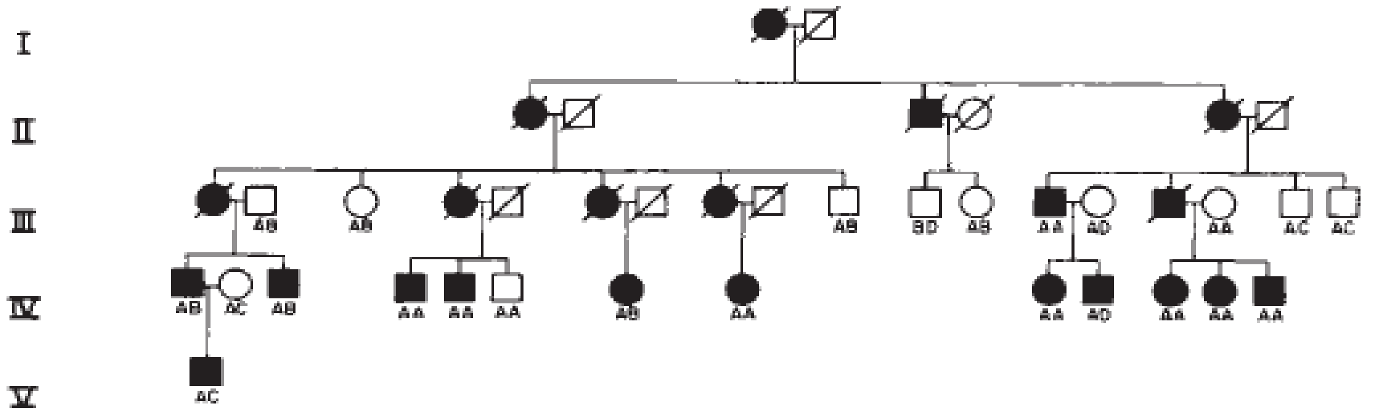
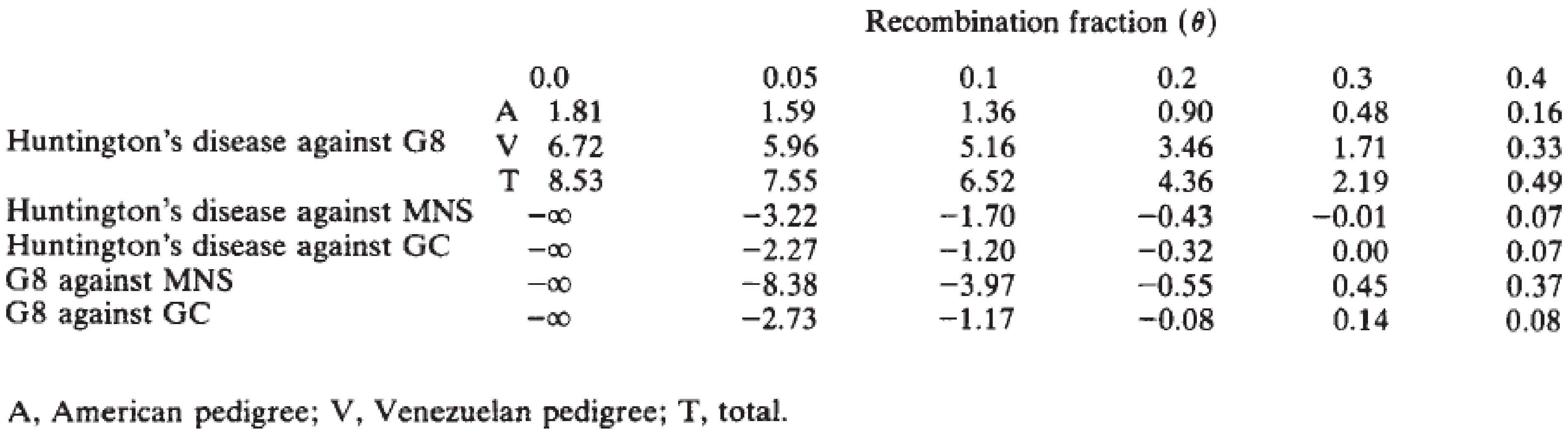


Fig. 1 Pedigree of an American Huntington's disease family. Symbols: circles, females; squares, males; a black symbol indicates that an individual is affected with Huntington's disease; a slashed symbol indicates that an individual is deceased. This pedigree was identified through the National Research Roster for Huntington's Disease Patients and Families at Indiana University. Relevant family members were examined by a neurologist and blood samples were obtained. EBV-transformed lymphoblastoid cell lines were established for the individuals whose genotypes are shown and have been stored at the Human Genetic Mutant Cell Repository, Camden, New Jersey. Phenotypes at the G8 locus shown under each symbol were determined by Southern blotting as outlined in Fig. 3. For the purposes of confidentiality, selected individuals are not shown.

Linkage of Huntington’s disease



		Recombination fraction (θ)					
		0.0	0.05	0.1	0.2	0.3	0.4
Huntington’s disease against G8	A	1.81	1.59	1.36	0.90	0.48	0.16
	V	6.72	5.96	5.16	3.46	1.71	0.33
	T	8.53	7.55	6.52	4.36	2.19	0.49
Huntington’s disease against MNS		−∞	−3.22	−1.70	−0.43	−0.01	0.07
Huntington’s disease against GC		−∞	−2.27	−1.20	−0.32	0.00	0.07
G8 against MNS		−∞	−8.38	−3.97	−0.55	0.45	0.37
G8 against GC		−∞	−2.73	−1.17	−0.08	0.14	0.08

A, American pedigree; V, Venezuelan pedigree; T, total.

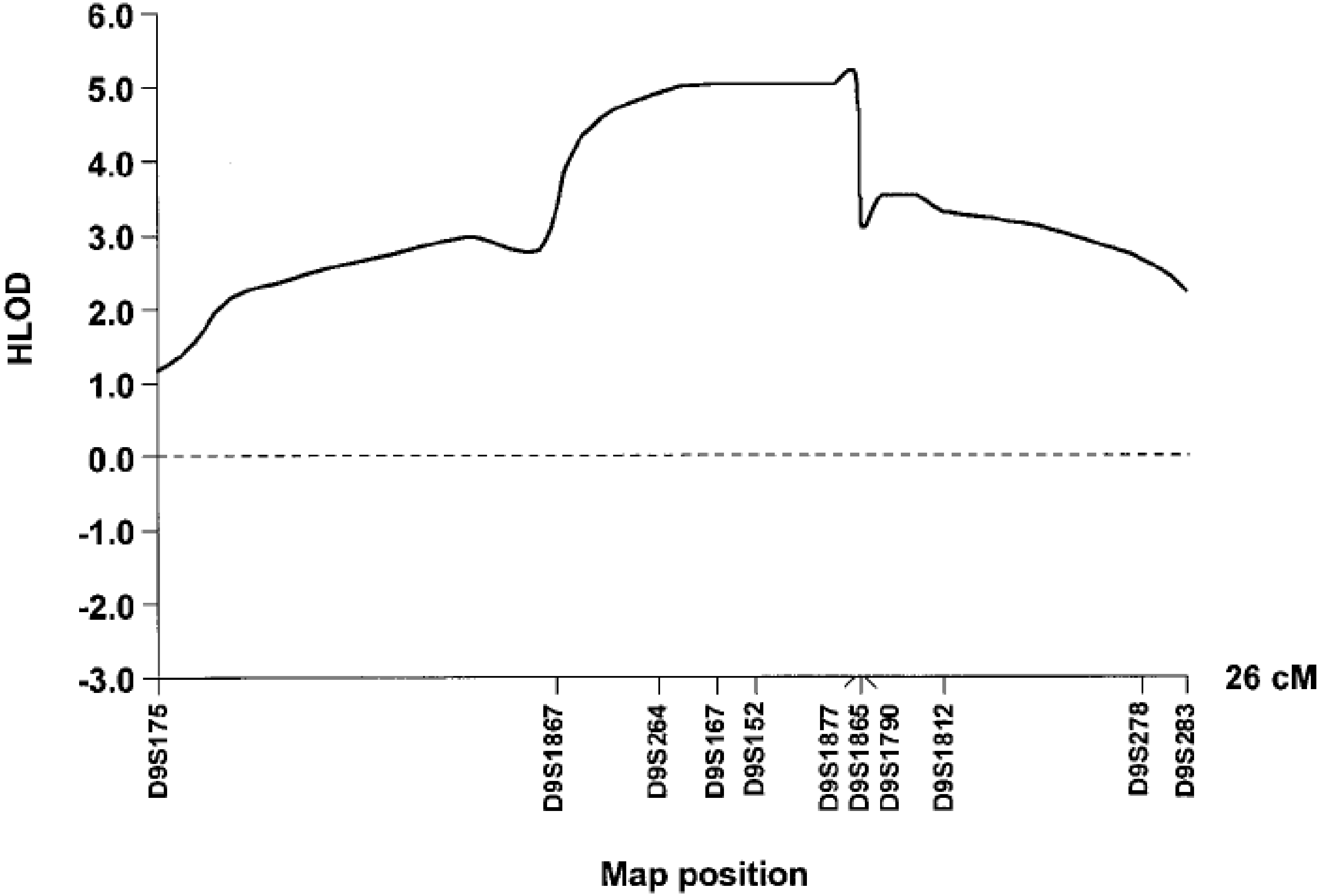
- The lod score between the Huntington’s disease locus and G8 locus at chromosome 4 is 6.52, with 99%CI 0-10cM. However there is no evidence of linkage with MNS and GC loci.

Familial hemophagocytic lymphohistiocytosis

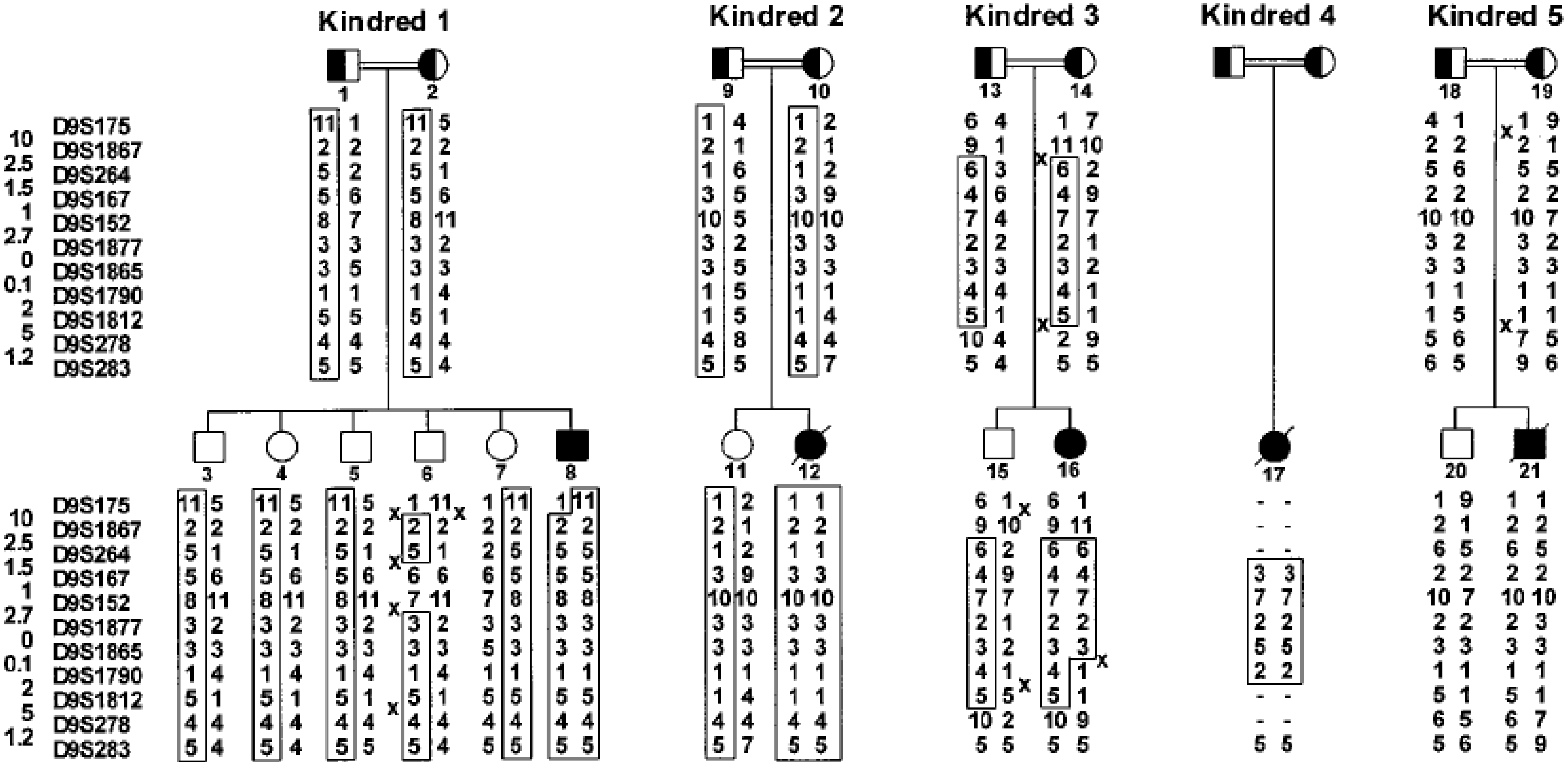
(Ohadi et al. 1999)

- An example of homozygosity mapping involving an autosomal recessive model with complete penetrance and a frequency of 0.004 for the disease allele
- Power was estimated by SLINK. Lod scores of 4.61 and 3.67 were obtained for markers 5 and 10 cM from the disease gene assuming five equi-frequent alleles.
- Genotyped done at the HGMP-RC.
- Allele frequencies were estimated from 50 unrelated, randomly selected, healthy Pakistani subjects. Analysis was done with GENEHUNTER. A MLOD of 4.40 and HLOD of 5.00 with α=0.81, at interval D9S1867-D9S1790. The admixture chi-squared test of heterogeneity gave a value of 2.76 and significant level 0.1.

HLOD of the five FHL families



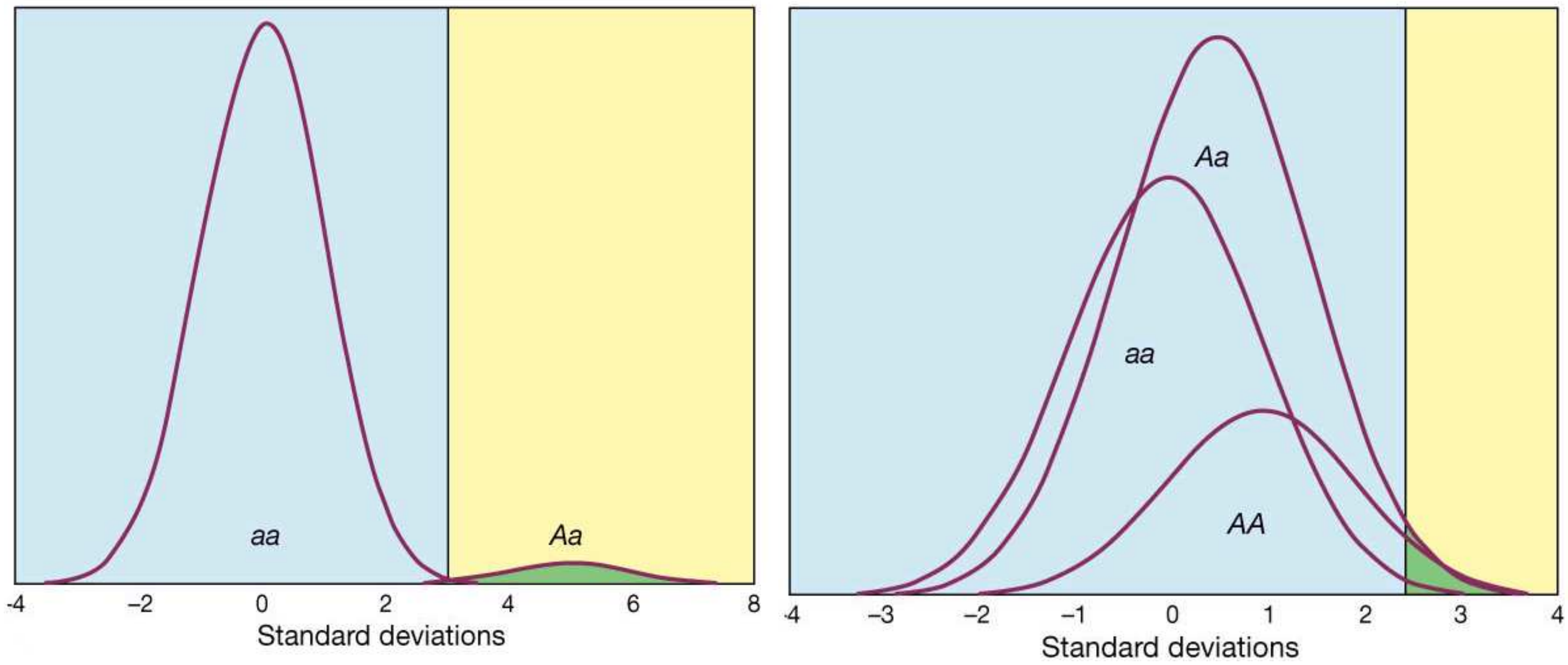
Allele segregation of linked and flanking markers



Genomewide significance level

- A lod score of 3, or a likelihood ratio of 1000, can be translated into the odds of ~20:1 in favor of linkage, or a significance level of 0.05.
- Let α_T=Pr(T>Z_{obs}|no linkage) at a specific position, α_T*=Pr(T>Z_{obs}|no linkage) anywhere in the genome.
- $\alpha_T^*=(C+9.2\rho GT)\alpha_T$, where T=threshold of lod score, C=the number of chromosomes (23), G=the total length of the genome in Morgans (33), ρ=crossover rate depending on the relationship.
  - For pedigree data, the genomewide significance of 0.05 is achieved at a pointwise significance of $4.9\text{x}10^{-5}$, or a lod score of ~3.3.
  - For ASP, the threshold is 3.6 with IBD testing and 4.0 for IBS testing.

Quantitative traits (Risch (2000) Nature)



- Mendelian (donminant) and non-Mendelian (additive) models of quantitative traits.

QTL analysis

- Haseman-Elston regression for sibling data
- Variance component model
- Regression conditional on trait
- Generalised estimating equations (GEE)

Haseman-Elston regression

- It detects linkage by evaluating the relationship between the similarity of trait values and the similarity of marker genotypes among relative pairs.
- A phenotype (x) is related to genotype (g) through a linear model $x=\mu+g+e$
- Application to the jth sib pair leading to $Y=(x_{1j}-x_{2j})^2$ and we consider all possible mating types to obtain $E(Y_j|\pi_j)=\alpha+\beta\ \pi_j$ at the trait locus. $\beta=-2\sigma_g^2$
- When markers are available, it has a similar form with the π_j being the estimated proportion of IBD at the marker locus, $\beta=-2(1-2\theta)^2\sigma_g^2$. Since $0\leq\theta\leq0.5$, a negative value of β indicate evidence of linkage.
- There have been numerous extension to this, and a notable one is the so-called conditional on trait method as implemented in Merlin-regress.

The New Haseman-Elston method

- It achieves generality by mean-centering trait values (z_j) for m members in the jth pedigree,

$$y_j = z_j - X_j\hat{\beta}$$

- Then we have the original HE regression by forming, vectorizing the (modified) trait cross-product, and equating the expected values.

$$E\begin{bmatrix} (y_{1k} - x_{1k}\beta)^2 \\ (y_{1k} - x_{1k}\beta)(y_{2k} - x_{2k}\beta) \\ \cdots \\ (y_{mk} - x_{mk}\beta)^2 \end{bmatrix} = \begin{bmatrix} \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{p}}^2 + \sigma_{\mathrm{c}}^2 + \sigma_{\mathrm{e}}^2 \\ \hat{\pi}_{12k}\sigma_{\mathrm{g}}^2 + 2\phi_{12k}\sigma_{\mathrm{p}}^2 + \sigma_{\mathrm{c}}^2 \\ \cdots \\ \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{p}}^2 + \sigma_{\mathrm{c}}^2 + \sigma_{\mathrm{e}}^2 \end{bmatrix}$$

Variance component analysis

- The study of continuous traits and their variation due to genetic and environmental factors.
- The variance of a quantitative trait is partitioned into various genetic and environmental components, including additive, dominance, shared and nonshared environment.
- The genetic component may be further partitioned into a QTL component and residual genetic component, when genetic marker data are available.
- The correlation of additive genetic component is determined by **kinship coefficient** (φ_{ij}, the probability of randomly drawing an allele in individual I being IBD to an allele at the same locus randomly drawn from individual j), the correlation of dominance is Δ_{7ij}, the probability that both alleles at the locus being IBD, the correlation of environmental factors is γ_{ij}.

Covariance of phenotype (X) between relatives i and j

$$Cov(X_i, X_j) = 2\phi_{ij}\sigma_a^2 + \Delta_{7ij}\sigma_d^2 + \gamma_{ij}\sigma_c^2 + \delta_{ij}\sigma_e^2$$

- an N x N expected variance-covariance matrix for the entire family of N individuals can be written and assumed to be from a multivariate normal distribution.
- The total likelihood is the product of all the likelihoods of individual families.
- Likelihood ratio test can be applied to nested hypotheses about the model parameters.
- Regardless of the dominance and environmental effects, the variance-covariance matrix consists of a term for estimated proportion of genes shared IBD at the QTL and genetic variance due to the QTL.

Power of variance components analysis

- For sib pairs, a test of the QTL variance has noncentrality parameter

$$\hat{\delta} = q^2(h^4 + 4)/2/(h^4 - 4)^2, \text{ where } q^2 = \sigma_q^2/\sigma^2,$$

$$h^2 = (\sigma_q^2 + \sigma_a^2)/\sigma^2,$$

- For a test of linkage having 80% power and size of 0.001 (a lod score of 3), the critical value of the chi-squared test statistic is 20.78, so that the number of sib pairs required is N = 20.78/$\hat{\delta}$, or 2N individuals.

Other aspects in linkage analysis

- Interference
- Imprinting
- Two-locus model
- Staged design

Software

- Crimap
- LIPED, LINKAGE, FASTLINK, MFLINK, VITESSE
- SLINK/FASTSLINK
- MENDEL, SIMLINK
- SUPERLINK
- MAPMAKER/GENEHUNTER, Allegro, MERLIN
- ASPEX, SPLINK
- MORGAN

To wrap up

- Linkage has played a central role in Mendelian disorders and has values for complex traits.
- There are several steps in conducting linkage studies, including pedigree ascertainment, study of power, genotyping, statistical analysis, and reporting.
- One of the major concerns of variance components model and Haseman-Elston regression methods is robustness, namely, the result of linkage analysis could depend on families with extreme scores.
- Linkage has limited use in fine-mapping and it is desirable to (1) redefine disease to increase the relative risk among relatives; (2) to fine-structure LD mapping using allelic or haplotype association.

LINKAGE

- It is available from course\linkage directory. It was developed by Lathrop, Ott et al. in the 1980s according to the Elston-Stewart algorithm which is appropriate for larger pedigrees with a few (multiallelic) markers.
- It consists of MLINK, ILINK, Makeped, LODSCORE, LINKMAP, UNKNOWN programs and utility programs such as lcp, lrp, lsp, preplink, loops.
- It requires data file (pedfile.dat) and parameter file (datafile.dat). They can be generated from utility programs and often used by other programs such as GENEHUNTER and SuperLink.
- Standard reference: Terwilliger & Ott (1994) Handbook of Human Genetic Linkage. The Johns Hopkins University Press.

Computer simulation programs

- SIMULATE: Pascal program to simulate pedigree data in a completely random fashion.
- SLINK: Pascal/C(FastSLINK) program to be able to condition on known marker information (Ott. PNAS 1989 **86**:4175-4178; Weeks *et al*. Am J Hum Genet 1990 **47:** A204; Cottingham et al. Am J Hum Genet 1993 **53**:252-263).
- SIMLINK: Fortran program similar to SLINK but the generated data has to be analyze under the same model (Boehnke. Am J Hum Genet 1986 **39**:513-527; Ploughman & Boehnke. Am J Hum Genet 1989 :543-551).
- CHRSIM: Pascal program possible to specify different map functions (Terwilliger *et al*. Genet Epidemiol 1993 **10:** 217-224).

Data files

- SIMULATE
  - Input: problem.dat, simped.dat, simdata.dat
  - Output: problem.dat, pedfile.dat, simout.dat
- SLINK
  - Input: slinkin.dat simped.dat, simdata.dat
  - Output: pedfile.dat, slinkin.dat
- CHRSIM
  - Input: simped.dat, input.dat
  - Output: pedfile.dat, test.res, outfile.dat
- SIMLINK: (omitted)
- Linkage analysis of pedigree files produced by SIMULATE and SLINK could be done with LINKAGE or SLINK (*msim, isim, lsim, elodhet*), *unknown* and possibly *lcp*/*lsp*/*lrp* are required in both cases. For SLINK, another file. limit.dat, is necessary

GENEHUNTER

- Implements parametric and nonparametric linkage analyses through Lander-Green algorithm. Usage:
- gh

-
  ```
  npl:1> photo sample
  npl:2> load linkloci.dat
  npl:3> scan linkped.pre
  npl:4> total stat het
  npl:5> quit
  ```

Exercise

- Please check course\linkage directory for the computer executables handdata.zip contains all the examples from the Terwilliger & Ott (1994) book, while ch_28.tar includes examples for computer simulation.

- Please also check the Rockefeller website for HTML documentation: http://linkage.rockefeller.edu and my personal page for exercise for the 1998 HGMP course.

Merlin

- A recent implementation of Lander-Green algorithm
- For more information, please check http://www.sph.umich.edu/csg/abecasis/Merlin/tour/linkage.html